

Finding Optimal Sinks for Random Walkers in a Network

Fern Y. Hunt

Applied and Computational Mathematics Division
National Institute of Standards and Technology
Gaithersburg, Maryland 20899

April 11, 2017

Abstract

In a model of network communication based on a random walk in an undirected graph, what subset of nodes (subject to constraints on the set size), enables the fastest spread of information? In this paper, we assume the dynamics of spread is described by a network consensus process, but to find the most effective seeds we consider the target set of a random walk—the process dual to network consensus spread. Thus an optimal set A minimizes the sum of the expected first hitting times $F(A)$, of random walks that start at nodes outside the set. Identifying such a set is a problem in combinatorial optimization that is probably NP hard. However F has been shown to be a supermodular and non-increasing set function and fortunately some results on optimization of such functions exist.

We introduce a submodular, non-decreasing rank function ρ , that permits some comparison between the solution obtained by the classical greedy algorithm and one obtained by our methods. The supermodularity and non-increasing properties of F are used to show that the rank of our solution is at least $(1 - \frac{1}{e})$ times the rank of the optimal set. When our approximation has a higher rank than the greedy solution, this can be improved to $(1 - \frac{1}{e})(1 + \chi)$ where $\chi > 0$ is a constant. A non-zero lower bound for χ can be obtained when the curvature and increments of ρ are known.

1 Introduction

The study of information spread (or dually consensus) in complex networks has been the subject of intense activity in the past decade ([35], [37], [26], [38] [3]) as reserachers study the role of distinguished subsets of nodes, such as “leaders” in consensus models and “influential spreaders” in models of information spread. In particular the research reported in references [38], [26], [3] developed methods for obtaining optimal spreaders as determined by some measure of subset performance.

Recall that at each time step of a network consensus process, the function value assigned to a node that is not a seed or leader is updated by averaging the current value with those of its neighbors according to a weighted Laplacian matrix which we take to be the transition matrix of a Markov chain. Seed nodes affect the information values of its neighbors during updating ,but the value assigned to the seed is unchanged. Over time the value for each node converges to a value dictated by the seed nodes. It is desired to select a set subject to cardinality constraints, that will maximize the rate of convergence. As discussed in [5], the rate of convergence is determined by the Perron-Frobenius eigenvalue of the averaging matrix restricted to the rows and columns indexed

by the set complement of the set of seed nodes. Thus an optimal seed set would minimize this eigenvalue. Rather than solve this problem directly we will seek instead the optimal sink (or target set) of the dual of the network consensus process: the associated random walk. For a set of nodes, we define $F(A)$ to be the sum of the expected first hitting times of random walks that start at nodes outside the set. As briefly mentioned in [5] the expected first hitting is related to the eigenvalue so that $F(A)$ is small if and only if the eigenvalue is small. From this point of view A is an optimal if $F(A)$ is minimized over all sets with the same or lesser cardinality. In the discussion that follows the set of all nodes in the network is denoted by V

1.1 Problem Background

Borkar, Nair and Sanketh, [5] showed that for subsets $A \subseteq B \subseteq V$ and $j \in V$, $F(A) - F(A \cup \{j\}) \geq F(B) - F(B \cup \{j\})$, that is, F is a supermodular function. Thus $-F$ is submodular so if it is bounded our problem is an instance of submodular maximization, a classic problem in combinatorial optimization. The problem for objective functions other than the one we discuss here has found wide application in many areas of computer science particularly machine learning (see e.g. the tutorials containing introductory material on submodular functions, [4], [7]). In 1987, Nemhauser, Wolsey and Fisher [34] proved the fundamental result that a set constructed by the greedy algorithm for maximizing a bounded submodular set function has an approximation ratio of $(1 - 1/e)$. Since then Borgs, Brautbar, Chayes and Lucier [3] and Sviridenko, Vondrak and Ward [41], showed that approximations of comparable or better quality could be obtained very efficiently using different methods. Recently, building on the foundational work of Fujishige [16], Bilmes, Iyer and Jegelka presented a framework that unifies many disparate approaches to the optimization theory by exploiting the convex-concave like properties of submodular functions [21]. Aspects of our method fit into this scheme as we discuss in Section 3. In [30], Mikesell, Kenter and Hicks following on [18] considered a number of heuristic approaches for solving the problem we discuss here for a number of graphs and compared it with our method. The results (see Figure 6, [30]) showed that it was no worse than the greedy method and was sometimes better. Understanding the reason for this was another motivation for this work.

There is a body of closely related work in the area of opinion dynamics where so-called stubborn agents can have a profound effect on the outcome of consensus algorithms, random gossip algorithms and similar decentralized protocols. See for example, [2], [36] and for work on the optimal placement of stubborn agents in the voter model of communication see [1]. Research on decentralized protocols appears to have originated with Borkar and Vairaya [6] and Tsitsiklis [42] who discussed communication in networks with nodes modeled by markov decision processes.

Results of our research are also relevant to the design of algorithms for routing in wireless communication systems when location information is not available [37], [23], identification of influential individuals in a social network [26] and in sensor placements for efficiently detecting intrusions in computer networks [29]. In their investigation of a leader-follower network, Clark, Bushnell and Poovendran [12], [13] demonstrated the connection between the rate of convergence to network consensus and a supermodular function closely related to ours. Furthermore they showed that the greedy approximation of the set of optimal leaders produces an approximation that is within $(1 - 1/e)$ of optimal.

Given a connected graph $G = (V, E)$ with N vertices V and edges E , information spreads through the network by a process that is dual to the direction of the random walk (see [31]). An optimal sink is defined in terms of a set function F where for a subset $A \subset V$, $F(A)$ is the sum of mean first arrival times to A by random walkers that start at nodes outside of A . If A is an effective target set for the random walks then $F(A)$ is small. Thus the optimal set (subject to a cardinality constraint

K) minimizes $F(A)$ subject to $|A| \leq K$,

$$\min_{A \subset V, |A| \leq K} F(A). \quad (1)$$

Recall that a random walker situated at a node $i \in V$, moves to a neighboring node $j \in V$ in a single discrete time step with probability,

$$\text{Prob}\{i \rightarrow j\} = \{p(i, j) > 0, \text{ if } (i, j) \in E, \quad p(i, j) = 0, \text{ if } (i, j) \notin E\} \quad (2)$$

NOTE: In this paper $p(i, j) = 1/\text{deg}(i)$ where $\text{deg}(i)$ is the degree of node i . However any transition probabilities for which the resulting Markov chain is ergodic can be used.

The matrix $\mathcal{P} = (p_{ij})_{i,j=1 \dots N}$ is the transition matrix of a Markov chain which in our choice or any choice of transition probabilities, is assumed to be irreducible and aperiodic ([25]). Starting at any node outside of A , a random walker first reaches the set A at a hitting time $T_A = \min\{n > 0 : X_n \in A\}$, where X_n is the node occupied by the walker at time n . The expected hitting time is $\mathbb{E}[T_A]$. If the walker starts at a fixed $i \notin A$, then the expected hitting time is the conditional expectation $\mathbb{E}[T_A | X_0 = i] = \mathbb{E}_i[T_A]$. Writing $h(i, A) = \mathbb{E}_i[T_A]$, the value of F at A is expressed as

$$F(A) = \sum_{i \notin A} h(i, A). \quad (3)$$

Given A , $F(A)$ can be evaluated by solving a suitable linear equation. Indeed a standard result in Markov chain theory [25] tells us that $h(i, A)$ is the i th component of the vector \mathbf{H} , which is the solution of the linear equation,

$$\mathbf{H} = \mathbf{1} + \mathcal{P}_A \mathbf{H}, \quad (4)$$

where $\mathbf{1}$ is a column vector of $N - |A|$ ones and \mathcal{P}_A is the matrix that results from crossing out the rows and columns of \mathcal{P} corresponding to the nodes of A . The value $F(A)$ is then the sum of the components of \mathbf{H} .

1.2 Our Contribution and Organization of the Paper

We present an approach to the solution of the optimization problem (1) that generalizes the classic greedy algorithm. In this paper we introduce a new constraint set (optimal and near optimal sets) that fits within the framework of classes of sets that have the property of being closed under addition and judicious deletion of elements. Such classes e.g. matroids and greedoids play an important role in the optimization of modular and sub(super) modular functions. Vertex covers are used to create optimal and near optimal sets and if high quality subsets exist, then the offered approximation is guaranteed to be better than the classic greedy algorithm. We present sufficient conditions for the performance ratio of our method to exceed the $(1 - \frac{1}{e})$ ratio obtained by Borkar et al, and by Nemhauser et al for the greedy algorithm. These results are to our knowledge new to the treatment of this problem and to the optimization of sub(super)modular functions generally.

The plan of the paper is as follows: section 2.2 contains a definition and discussion of optimal and near optimal sets ranked relative to a vertex cover of the graph G with cardinality C . In sections 2.1 and 2.2, we demonstrate how the method is applied to a graph using a collection of sets \mathbf{S} that are subsets of a vertex cover. If every vertex cover contained optimal sets as subsets, it would make sense to use this choice consistently. Unfortunately, optimality of a set is generally not preserved by the addition or deletion of nodes. We remedy this situation in part by introducing greedoids, a class of sets that are closed under the judicious deletion and addition of single elements.

$$\begin{array}{l}
\mathbf{S} \\
\{3,7\} \rightarrow \{3,5,7\} \rightarrow \{3,5,7,8\} \rightarrow \{1,3,5,7,8\} \\
\{1,7\} \rightarrow \{1,5,7\} \rightarrow \{1,5,7,8\} \rightarrow \{1,3,5,7,8\} \\
\{3,8\} \rightarrow \{3,6,8\} \rightarrow \{3,4,6,8\} \rightarrow \{2,3,4,6,8\} \\
\{3,6\} \rightarrow \{3,6,8\} \rightarrow \{3,4,6,8\} \rightarrow \{1,3,4,6,8\}
\end{array}$$

Figure 1: Optimal sets of the graph in Figure 2 for $K=4$ and 5 obtained by greedy extension of \mathbf{S} , subsets of a vertex cover

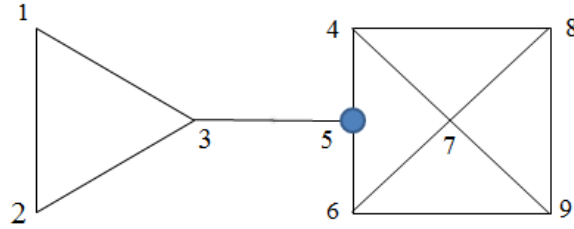


Figure 2: Graph with $N=9$ vertices, showing optimal set for $K=1$

In section 3 we demonstrate the method on a second graph where \mathbf{S} is chosen to be a group of feasible sets of a greedoid. In section 4, the quality of the approximation obtained by our method is evaluated in terms of the ranking function $\bar{\rho}$ introduced in section 2.2. A discussion of the computational complexity and tradeoff considerations can be found in section 4.2. The statement and proof of the main result we described earlier is in section 4.3. Our concluding discussion is found in section 5.

2 Finding and Approximating Optimal Sets

2.1 Maximal Matches

The optimization problem as posed in equation (1) assumes no advance knowledge about the optimal set or any other possibly related sets. Thus let us first consider a process of obtaining optimal sets by using subsets of existing ones. Let A be a vertex cover (not necessarily a minimum one). Since every edge is incident to a vertex of A , a random walker starting at a vertex i outside of A must hit A at the first step. That is $h(i, A) = 1$. Now equation (4) implies that $h(i, A) \geq 1$ so it follows that A must be an optimal set for its own cardinality. Thus a solution for $C = |A|$ is obtained by constructing a vertex cover. Fortunately a maximal match can be constructed by a simple greedy algorithm and its vertices are a vertex cover with cardinality $C \leq 2*(\mathcal{VC})$ where \mathcal{VC} is the cardinality of a minimum vertex cover [14]. Therefore without loss of generality we turn our attention to the solution of problem (1) for $K \leq C$.

2.2 Optimal and Near Optimal Sets

We introduced a measure of the spread effectiveness of sets in equation (3). It will be convenient to convert this to a rank defined on subsets of V . In particular, suppose there exists a vertex cover with C vertices. We will order all non-empty subsets $A \subseteq V$ such that $|A| \leq C$ with a ranking function $\bar{\rho}(A)$ defined as,

$$\bar{\rho}(A) = \frac{F_{max} - F(A)}{F_{max} - F_{min}} \quad (5)$$

where $F_{max} = \max_{\emptyset \neq A \subseteq V, |A| \leq C} F(A)$, and F_{min} is the corresponding minimum. F_{min} can be calculated by computing F for a vertex cover of cardinality C whose elements might be e.g. the endpoints of a maximal match. We define F_{max} to be the maximum value of F among all one element subsets. We assume that $F_{max} \neq F_{min}$. If this were not the case, $F(A)$ would have the same value for any non-empty subset A with $|A| \leq C$. Thus any A with $|A| \leq K$ would be a solution of the problem.

If A is optimal and $|A| = C$ then $\bar{\rho}(A) = 1$. Conversely, the worst performing set has value 0. For a constant ν ($0 < \nu \leq 1$) and C , the non-empty set

$$L_{\nu,C} = \{A : A \subseteq V, |A| \leq C, \bar{\rho}(A) \geq \nu\} \quad (6)$$

defines a set of optimal and near optimal subsets, with the degree of near optimality depending on ν . Let m be the smallest cardinality of sets in $L_{\nu,C}$. Starting with a collection of sets $\mathbf{S} \subset L_{\nu,C}$ of size m , our method is to seek a solution to the optimization problem (1) by greedily augmenting each set until it reaches the desired size K . The offered approximation is the best (has the lowest F value) of these extended sets. We can always find a ν and C so that $L_{\nu,C}$ contains the optimal set of cardinality K but we do not have a proof that the approximation generated by subsets of a vertex cover is optimal. However since our solution is a superset of sets in $L_{\nu,C}$, it is also in $L_{\nu,C}$ and therefore has minimum rank ν . We illustrate the method with an example.

Figures 2 through 6 show a graph with $N = 9$ vertices along with the vertices of optimal sets for $K = 1$ through 5. To solve the problem for $K = 4$, we note that the class of optimal and near optimal sets based on $C = 8$ and $\nu = 0.90$ has minimum set size $m = 2$. The set $\mathcal{M} = \{1, 3, 5, 6, 7, 8\}$ is a vertex cover (calculated from the maximal match algorithm). We define \mathbf{S} to be the two element subsets of \mathcal{M} that are in $L_{.90,8}$. The first column of Figure 1 lists these sets and subsequent columns show the results of greedy one element extensions of \mathbf{S} until $K = 4$. Optimal sets are shown in red. In this example the offered approximation is optimal. This is also the case for extensions up to $K = 5$. In this case we see that the method identifies optimal sets that are subsets of \mathcal{M} as well as others that are not, e.g., $\{2, 3, 4, 6, 8\}$, underlining the fact that the method finds sets that are reachable by greedy extension of subsets of \mathcal{M} . The offered approximation for this method is guaranteed to be in $L_{.90,8}$. This is a consequence of Proposition 1 which is discussed and proved in Section 3.

3 Closure Property of Optimal and Near Optimal Sets

In section 2.2, we demonstrated our method for approximating a solution of optimization problem (1) based on greedy extensions of subsets of a vertex cover that are optimal or near optimal. Unfortunately a vertex cover can fail to have such subsets other than the vertex cover itself (see an example in [18]). This is the motivation for finding other classes of optimal and near optimal sets that are closed under judicious addition and deletion of elements. We conjecture that greedy extension of such sets will have the largest likelihood of success. The structure we seek is conveniently described in terms of a generalization of the matroid, known as a *greedoid* [28, 8].

Definition 1 Let \mathbf{E} be a set and let \mathcal{F} be a collection of subsets of \mathbf{E} . The pair $(\mathbf{E}, \mathcal{F})$ is called a greedoid if \mathcal{F} satisfies

- **G1** : $\emptyset \in \mathcal{F}$
- **G2** : For $A \in \mathcal{F}$ non-empty, there exists an $a \in A$ such that $A \setminus \{a\} \in \mathcal{F}$
- **G3** : Given $X, Y \in \mathcal{F}$ with $|X| > |Y|$, there exists an $x \in X \setminus Y$, such that $Y \cup \{x\} \in \mathcal{F}$

A set in \mathcal{F} is called feasible. Note that **G2** implies that a single element can be removed from a feasible set X so that the reduced set is still feasible. By repeating this process the empty set eventually is reached. Conversely starting from the empty set, X can be built up in steps through repeated use of **G3**.

We now show that $L_{\nu, C}$ satisfies condition **G3** of the definition for any $0 < \nu \leq 1$, $0 \leq C \leq N$ (Proposition 1). The proof depends on the following lemma and uses an adaptation of an argument in Clark et al [12]

Lemma 1 Let $S \subseteq V$, $u \in V \setminus S$. Then $F(S) \geq F(S \cup \{u\})$.

Proof: Suppose S -a set of nodes, is a target set for a random walk. Let $E_{ij}^l(S)$ be the event, $E_{ij}^l(S) = \{X_0 = i \in V, X_l = j \in V \setminus S, X_r \notin S, 0 \leq r \leq l\}$. Thus paths of the random walk in this event start at i and arrive at j without visiting S during the interval $[0, l]$. Also define the event $F_{ij}^l(S, u) = E_{ij}^l(S) \cap \bigcup_{m=0}^l \{X(m) = u\}$ where $u \notin S$. Paths in this event also start at i and arrive at j without visiting S , but must visit the element u at some time during the interval $[0, l]$. Since a path either visits u in the time interval $[0, l]$ or it does not, it follows that:

$$E_{ij}^l(S) = E_{ij}^l(S \cup \{u\}) \cup F_{ij}^l(S, u) \quad (7)$$

We have $E_{ij}^l(S \cup \{u\}) \cap F_{ij}^l(S, u) = \emptyset$. This implies that,

$$\mathbf{1}_{E_{ij}^l(S)} = \mathbf{1}_{E_{ij}^l(S \cup \{u\})} + \mathbf{1}_{F_{ij}^l(S, u)} \quad (8)$$

and therefore:

$$\mathbf{1}_{E_{ij}^l(S)} \geq \mathbf{1}_{E_{ij}^l(S \cup \{u\})} \quad (9)$$

Here $\mathbf{1}_A$ is the usual indicator function of the set A , i.e. the function $\mathbf{1}_A : \Omega \rightarrow \{0, 1\}$. Recalling that T_S is the hitting time for set S , the following relation comes from taking the expectation of $\mathbf{1}_{E_{ij}^l(S)}$ on the left hand side of (9) summing over all $j \in V \setminus S$. Here \mathbb{E} denotes expectation.

$$\mathbf{Prob}\{T_S > l | X_0 = i\} = \mathbb{E} \left(\sum_{j \in V \setminus S} \mathbf{1}_{E_{ij}^l(S)} \right) \quad (10)$$

A similar result is obtained for $T_{S \cup \{u\}}$ from taking the expectation of $\mathbf{1}_{E_{ij}^l(S \cup \{u\})}$ on the right hand side of (9) and summing over $j \in V \setminus S$. Summing once again over all $l \geq 1$ results in the inequality,

$$h(i, S) \geq h(i, S \cup \{u\}) \quad (11)$$

□

REMARK: See (section 5 in [19]), for an explicit formula for the increment $F(S) - F(S \cup \{u\})$.

Proposition 1 For $0 < \nu \leq 1$ and $0 < C \leq N$, let $L_{\nu,C}$ be the class of sets defined in equation (6). Then $L_{\nu,C}$ satisfies condition **G3**.

Proof: The conclusion follows from the definition of $L_{\nu,C}$ and Lemma 1. \square

The proposition establishes that $L_{\nu,C}$ satisfies the **G3** property for greedoids. However, **G2** does not hold. For example if the set A has cardinality m where m is the size of the smallest set in $L_{\nu,C}$ then $A \setminus \{a\}$ cannot be in $L_{\nu,C}$ for any element $a \in A$. Conversely, let $c_n = \max_{|X| \leq n} \rho(X)$. If $c_m \geq c > c_{m-1}$ then m is the size of the smallest set in $L_{\nu,C}$. Define G_n to be all sets in $L_{\nu,C}$ of cardinality n . To create a class of sets with the **G2** property, one constructs subsets of G_m of size $n \leq m$ that are "augmentable", i.e., that satisfy **G3**. Sets G_n for $n > m$ are culled so the remaining sets are supersets of the "augmentable" sets and therefore satisfy **G2**. The greedoid will then consist of selected subsets and supersets of G_m . Conditions for the existence of "augmentable" subsets of G_m and proof of the validity of the resulting greedoid construction can be found in [18]. Rather than repeat the details of these arguments here, we close this section with an example showing the greedoid of a graph (Figure 7) and its use in the solution of (1). The minimum cardinality of a set in the class of optimal and near optimal sets $L_{.85,7}$ is $m = 3$. These sets are used to create the greedoid depicted in Figure 8. Note that **G1-G3** are satisfied. Assume the optimal set for $K = 4$ is unknown. Then our method in this case is to take **S** to be the three element sets in $L_{.85,7}$ that are feasible sets of the greedoid and perform a greedy extension of each set. In Figure 8 a line is drawn between a set and its greedy extension. We have also drawn greedy extensions of sets of cardinality $n < m$ as well. The optimal sets are shown in red and so they are in the greedoid. The offered approximations are in fact exact. This second general approach of using the starter set **S** to be a collection of feasible sets of a greedoid contains the vector cover based approach as a special case. Note that the vector cover subsets in the starter set are restricted to be optimal and near optimal sets in $L_{\nu,C}$ rather than arbitrary subsets. Thus they must have a minimum quality ν and have cardinality less than C .

The methods we present here are similar to some earlier approaches to the problem. For example the enumeration technique for maximizing a submodular function subject to a modular function constraint ([39],[27], [40]). Rather than begin the greedy algorithm with an optimal singleton, the enumeration technique performs a greedy extension of all k element subsets (where $k \leq m = 3$). Khuller et al ([27]) and then Sviridenko ([40]) proved these methods have a $(1 - \frac{1}{e})$ performance guarantee and Feige in ([15]) proved this guarantee was the best possible. In our setting, m is associated with a lower bound of the rank of near optimal sets ν . In section 4.3 we show how an improvement in the performance guarantee is possible if the rank of the greedy solution is less than some η where $\eta \geq \nu$. Our method also has connections to the MMax the ascent algorithm for constrained maximization of a monotone submodular function, applied to $-F$ (see section 6.2 in [21]). We iterate a greedy subgradient equation for each member of the starting set **S**. Successive iterates are non-decreasing and converge to an extreme point of the subdifferential of $-F$ at the K element terminal set containing the initial set([22]).

4 Quality of the approximation

4.1 Comparison with the optimal solution and greedy solution

Following Ilev ([20]), F can be defined for the empty set as

$$0 \leq F(\emptyset) = \max_{X \cap Y = \emptyset, X, Y \subseteq V} F(X) + F(Y) - F(X \cup Y) < \infty \quad (12)$$

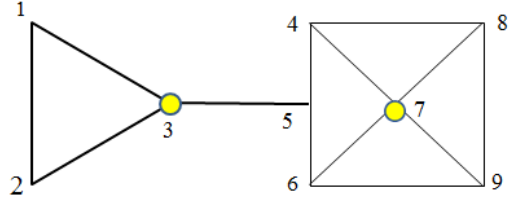


Figure 3: Optimal set $K=2$ for graph in Figure 2

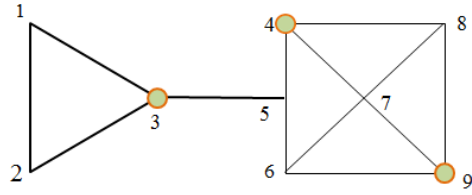


Figure 4: Optimal set $K=3$ for graph in Figure 2. The set $\{3,6,8\}$ is also optimal by symmetry

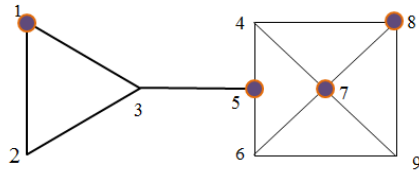


Figure 5: Optimal set $K=4$ for graph in Figure 2

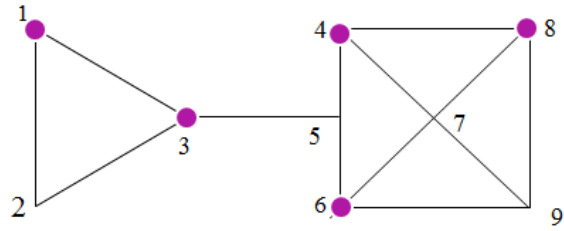


Figure 6: Optimal set $K=5$ for graph in Figure 2

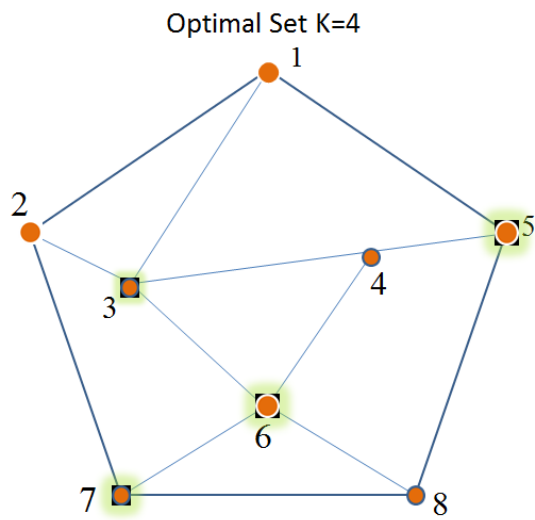


Figure 7: Graph with $N=8$, vertices. Vertices of optimal set $K=4$ shown as squares

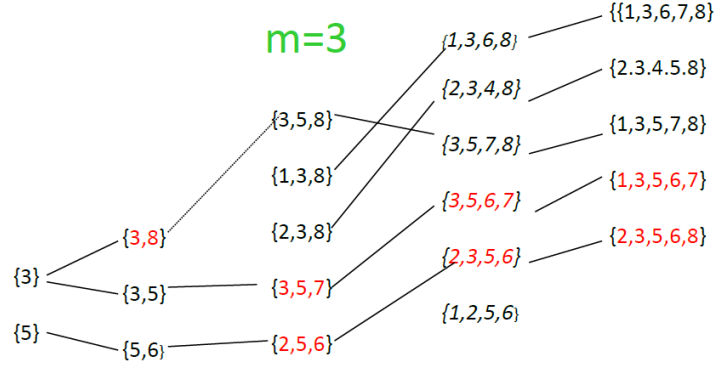


Figure 8: Greedoid constructed from optimal and near optimal sets $L_{85,7}$ of graph in Fig 7

Thus by the definition of $\bar{\rho}$, $\bar{\rho}(\emptyset) = \frac{F_{max} - F(\emptyset)}{F_{max} - F_{min}}$. This means the normalized function defined on sets A , $\rho(A) = \bar{\rho}(A) - \bar{\rho}(\emptyset)$ is bounded, submodular and non-decreasing. Also note that since $F(\emptyset) \geq F_{max}$, $\rho(A) \geq 0$ for all $A \subseteq V$. For the empty set we have $\rho(\emptyset) = 0$. For the remainder of this paper, we will refer to ρ as the normalized rank function and to $\bar{\rho}$ as the un-normalized rank function. Since the rank is an affine function of F , the optimization problem (1), is equivalent to the problem of finding the set that maximizes the normalized rank subject to the same constraints. This is a special case of the general problem first considered by Nemhauser, Wolsey and Fisher [34].

Our offered solution is the result of the greedy extension of a group of optimal and near optimal sets of minimal cardinality m . In this section we will compare the solution to the optimal solution using the normalized rank function ρ . Specifically, we show that the inequality of Nemhauser, Wolsey and Fisher, ([34], Section 4) holds. Recall that they proved that the ratio of the normalized rank of the classic greedy solution to the optimal solution has a lower bound of $(1 - \frac{1}{e})$.

Let $S_g^{(m)}$ be the m element set that is the result of greedily adding single elements m times. We first suppose that $S_g^{(m)} \in \mathbf{S}$.

Lemma 2 Suppose $S_g^{(m)} \in \mathbf{S} \subseteq L_{\nu,C}$. Let S_g be the K element set obtained from the greedy extension of S_g . If S^* is the offered solution, then

$$F(S^*) \leq F(S_g) \quad (13)$$

Proof: $F(S^*)$ is the minimum value of all the values obtained by the greedy $K - m$ extension of elements in \mathbf{S} . \square

The set S_g is also the result of greedily adding single elements K times. Thus we may use section 4 in [34] and the definition of ρ to conclude that

Corollary 1 If S^* is the solution constructed by the method described in sections 2.2 and 3 above, then

$$\rho(S^*) \geq (1 - \frac{1}{e})\rho(\mathcal{O}_K^*) \quad (14)$$

where \mathcal{O}_K^* is the solution of the optimization problem.

Once $F(S^*)$ and $F(S_g)$ have been computed we can determine χ such that $\rho(S^*) = (1 + \chi)\rho(S_g)$. The quantity χ measures the degree of improvement of $F(S^*)$ over $F(S_g)$. In particular we have,

Proposition 2 When $F(S^*) < F(S_g)$, so $\chi > 0$, then

$$\rho(S^*) \geq (1 + \chi)(1 - \frac{1}{e})\rho(\mathcal{O}_K^*) \quad (15)$$

NOTE: If $m = 1$, then optimal and near optimal singletons of quality ν would be in $L_{\nu,C}$. Thus necessarily $S_g^{(m)}$ must be in \mathbf{S} . In other words for $m = 1$, our method is a direct generalization of the classic greedy method.

If $S_g^{(m)} \notin \mathbf{S}$, then it may be possible to identify a set S with $\rho(S) > \rho(S_g)$ and $|S| \leq K$. In section 4.3 we discuss how this can be done when bounds on the curvature of ρ and increments of F are available. The existence of such an S allows us to formulate sufficient conditions for S^* to satisfy the hypothesis of Proposition 2.

A lower bound of $(1 - \frac{1}{e})$ was established by Borkar et al in [5]. Specifically it is a lower bound on the ratio of $F(S_g) - F(\{a\})$ to $F(\mathcal{O}_K^*) - F(\{a\})$, where S_g is the result of the greedy algorithm starting with singleton a .

4.2 Computational effort and tradeoff with quality

A rough estimate of the complexity of the method follows from realizing that the collection $\mathbf{S} \in L_{\nu,C}$, has at most $\binom{N}{m}$, m element sets. To determine whether or not a particular set is near optimal, equation(4) must be solved and this involves $O(N^3)$ operations. Thus the objective function values of \mathbf{S} are determined in $O(N^{m+3})$ operations. The greedy extension of an m element to a K element set involves $O((K-m)(N-m)) = O(N^2)$ so that the extension of every set in \mathbf{S} involves $O(N^{m+2})$ operations. Overall then, the method requires $O(N^{m+3}) + O(N^{m+2}) = O(N^{m+3})$ operations. Thus it is desirable to make m as small as possible. In fact we assume $m \ll K$. However the size of m affects the accuracy. Taking ν to be a measure of the quality of the approximation, we want to know given m , what ν can be expected? Conversely given a desired quality ν , what m is required? We will employ the forward elemental curvature of the normalized rank function. Elemental curvature was used by Wang, Moran, Wang, and Pan [43] in their treatment of the problem of maximizing a monotone non-decreasing submodular function subject to a matroid constraint.

The elemental curvature of ρ is defined over $L_{\nu,C}$ in terms of the marginal increase in the rank of a set when a single element is added to it. First let A be a set and $i \notin A$,

$$\rho_i(A) = \rho(A \cup i) - \rho(A). \quad (16)$$

and then for a fixed $A \in L_{\nu,C}$ set,

$$k_{ij}(A) = \frac{\rho_i(A \cup j)}{\rho_i(A)}. \quad (17)$$

The curvature is defined then as,

$$\kappa = \max\{k_{ij}(A) : A \in L_{\nu,C}, i \neq j, i, j \notin A\}. \quad (18)$$

Since ρ is submodular $\kappa \leq 1$. Now suppose $S \subset T \subset L_{\nu,C}$. Given ν , we want to determine the minimum size of S for which $\rho(S) \geq \nu$. If $T \setminus S = \{j_1, \dots, j_r\}$, we have (see equation (2) in [43]),

$$\bar{\rho}(T) - \bar{\rho}(S) = \rho(T) - \rho(S) = \sum_{t=1}^r \rho_{j_t}(S \cup \{j_1, \dots, j_{t-1}\}). \quad (19)$$

Therefore ,

$$\bar{\rho}(T) - \bar{\rho}(S) \leq \rho_{j_1}(S) + \kappa \rho_{j_2}(S) + \cdots \kappa^{t-1} \rho_{j_r}(S) \quad (20)$$

Suppose $\bar{\rho}(T) = 1$, for example if T is a vertex cover. Define γ to be $\gamma = \max\{\rho_{j_t}(S) : S \subset T, t = 1 \cdots r\}$. We can get a lower bound on the rank of S using equation (20) and the inequality $0 \leq \rho_j(S) \leq \gamma$. First assume γ is known. We know that if $S \neq \emptyset$, then $\gamma < 1$. Then,

$$\bar{\rho}(S) \geq 1 - \gamma \sum_{t=1}^r \kappa^{t-1} \quad (21)$$

Let us now suppose that :

$$\bar{\rho}(S) \geq \bar{\eta} = (1 - \gamma \sum_{t=1}^r \kappa^{t-1}) \geq \nu, \quad (22)$$

and $|S| \geq m$. If an approximation with quality ν is required, and $r(\nu)$ is the largest value of r such that inequality (22) holds, then $r \leq r(\nu)$. Now $K = C - r$ is the cardinality of S so that $C - r(\nu) \leq C - r$. Thus the smallest possible value of $|S|$ is

$$m(\nu) = C - r(\nu) \quad (23)$$

In particular any m must satisfy $m \geq m(\nu)$. Conversely, given m , the quality of the approximation depends on γ and $r = C - m$. More precisely, the largest value of ν and thus the largest guaranteed quality of an approximation obtained by our method, has an upper bound given by the right hand side of inequality (21).

4.3 When is our approximation is better than the greedy solution?

Clearly if $F(S^*)$ is available then we can check that $F(S^*) < F(S_g)$. However the value of $F(S^*)$ is needed to calculate the degree of improvement. If additional knowledge of the graph, expressed as information about F is available an a priori estimate of the degree of improvement is possible without calculating $F(S^*)$. When T is a vertex cover we obtained a lower bound on the rank of $S \subseteq T$ in terms of upper bounds on the increments and curvature of ρ and the rank of T . With this bound in hand we would like to sharpen the comparison between the rank of the set approximation S^* obtained by our method and the rank of the optimal set. We will derive a lower bound on the constant χ . To do this, notice that if the un-normalized rank of S^* satisfies $\bar{\rho}(S^*) > \bar{\eta}$ and the greedy solution for the same cardinality satisfies $\bar{\rho}(S_g) \leq \bar{\eta}$, the normalized ranks of the respective sets satisfy $\rho(S^*) > \eta$, and $\rho(S_g) \leq \eta$, where $\eta = \bar{\eta} - \bar{\rho}(\emptyset)$. Now as in section 4.1, $\rho(S^*) = (1 + \chi)\rho(S_g)$. So if we write $\rho(S_g) = (1 - \delta)\eta$ for some $0 < \delta < 1$, then $\rho(S^*) > \eta$ implies that $1 + \chi > \frac{1}{1 - \delta}$. Thus $\chi > \frac{\delta}{1 - \delta}$.

Rather than S^* we could easily suppose there exists some S that is a K element greedy extension of a set in \mathbf{S} such that $\rho(S) > \eta \geq \nu$. Arguing as before can conclude that $\rho(S) = (1 + \chi)\rho(S_g)$ for some $\chi > \frac{\delta}{1 - \delta}$. Since $\rho(S^*) \geq \rho(S)$, then we can still conclude that χ in inequality(15) has the claimed lower bound. Thus in summary, information about the the curvature and increments of $\bar{\rho}$ (equivalently ρ), can be used to obtain a lower bound $\eta > 0$ on the rank of a set S of cardinality K when $S \subseteq T$, a vertex cover. Let S also be the greedy extension of an element in \mathbf{S} , in other words, it is a set that arises from the implementation of our method. For example \mathbf{S} could be collection of subsets of a vertex cover T . Alternatively, \mathbf{S} could be a more general class of m element feasible sets of a greedoid. Greedy extensions of such sets can be subsets of a vertex cover (see the implementation for the graphs in this paper for example). If $\rho(S_g) < \eta$ then inequality (15) holds with $\chi > \frac{\delta}{1 - \delta}$.

Thus we have a direct comparison of our approximation S^* with the optimal solution. The degree of improvement over the classic greedy method, i.e. χ depends on the greedy solution S_g , η , ν and T . This is the content of the following proposition.

Proposition 3 *Assume the following conditions:*

1. *Let T be a vertex cover and for $0 < \nu < 1$, let $\mathbf{L}_{\nu,C}$ be a corresponding class of optimal and near optimal sets. Let $S \subseteq T$ be a K element set that is the greedy extension of a set in $\mathbf{S} \subset \mathbf{L}_{\nu,C}$.*
2. *$\rho(S) > \eta > 0$ where $\eta = \bar{\eta} - \bar{\rho}(\emptyset)$ and $\bar{\eta}$ is defined in inequality (22).*
3. *S_g is a K element set resulting from the greedy method. The rank of S_g satisfies, $\rho(S_g) < \min(\nu, \eta)$*

Then the offered approximation S^ satisfies*

$$\rho(S^*) \geq (1 + \chi)(1 - 1/e)\rho(\mathcal{O}_K^*) \quad (24)$$

where \mathcal{O}_K^ is the optimal solution of problem (1) and for some δ , $0 < \delta < 1$, χ satisfies,*

$$\chi > \frac{\delta}{(1 - \delta)}. \quad (25)$$

We may set $1 - \delta = \frac{\rho(S_g)}{\eta}$ so that δ depends on S_g , η , ν and T .

REMARK: Our discussion suggests S might be found by applying the backward greedy (worst-out) algorithm to a vertex cover with $T : |T| > K$ until a set of cardinality K is reached. The conditions of the proposition can then be checked on the resulting S . If S has a higher rank than the greedy solution but is not the greedy extension of a set in \mathbf{S} , then a better approximation or solution of the problem may result from a series of comparisons or single element exchanges between S^* and S .

In this section we stated conditions that imply the existence of a sets S whose rank (as expressed in terms of the elemental forward curvature of ρ), exceeds the rank of the greedy solution. If S is also a greedy extension of a set in \mathbf{S} then S^* our offered solution has a higher degree of optimality as expressed in inequalities (24) and (25)

5 Conclusion

We posed the problem of identifying a subset of nodes in a network that will enable the fastest spread of consensus in a decentralized communication environment. In a model of communication based on a random walk on an undirected graph $G = (V, E)$, the optimal set of nodes is found by minimizing the sum of the mean times of first arrival to the set by walkers who start at nodes outside the set. Since the objective function for this problem is supermodular, the greedy algorithm has been a principal method for constructing approximations to optimal sets. Previous results guarantee that these sets are in some sense within $(1 - 1/e)$ of optimality.

In this work we took a different approach. Rather than consider the problem (1) over all subsets of cardinality up to K , we restricted the search for an optimizing set to classes of optimal and near optimal sets that are closed under the addition and judicious deletion of elements. These sets have a predefined degree of near optimality (see section 2.2). Let m be the minimum cardinality of sets in this class. We offered an approximation of the solution of optimization problem based on the

greedy extension of a starter set \mathbf{S} of sets of size m . In actual implementation we take $m \leq 3$ and theoretical arguments suggest this is a good choice. We demonstrated the method for two choices of \mathbf{S} , first a class subsets of a vertex cover and then for the feasible sets of a greedoid constructed from $L_{\nu,C}$. Here ν measures the degree of near optimality relative to a vertex cover of cardinality C . As shown in section 4.1, when the greedy solution at stage m is in the initial set \mathbf{S} then the results of our method are (unsurprisingly) at least as good. It is also clear that if \mathbf{S} has a greedy extension that contains a set $S, |S| \leq K$, that is better than the greedy solution of problem (1), our offered approximation S^* must also be better. Our principal result is a set of sufficient conditions and a proof that they imply the existence of such a set (see section 4.3). It is a high ranking subset of a vertex cover and its rank can be estimated when information about the elemental forward curvature and increments of F are available. Moreover S^* can be directly compared with optimal solution and a lower bound on the improvement in optimality can be obtained without explicit knowledge of S^* .

In closing, we propose that the method we describe here is an instance of a more general approach based on the navigation in a graph whose nodes are optimal and near optimal sets. In this paper, the search for an optimal solution is based on greedily moving forward. However the most effective approach may be a more general class of steps. Movement between adjacent nodes would correspond to the addition (forward) and deletion (backward) or exchange of single elements.

References

- [1] D.Acemoglu, G.Como, F. Fagnani, A.Ozdaglar, *Opinion fluctuations and disagreements in social networks*, Mathematics of Operations Research, 201338:1, pp.1-27
- [2] W.Ameur, P. Bianchi, J. Jakubowicz, *Robust Average Consensus Using Total Variation Gossip Algorithm*, 2012 6th International ICST Conference on Performance Evaluation Methodologies and Tools (VALUETOOLS), Cagse, pp. 99-106
- [3] C. Borgs, M. Brautbar, J. Chayes, B. Lucier, *Maximizing Social Influence in Nearly Optimal Time* Proceedings of the 25th Annual ACM-SIAM Symposium on Discrete Algorithms, 2014, pp. 946-957
- [4] F.Bach, *Machine learning with Submodular Functions: A Convex Optimization Perspective*, <http://arxiv.org/abs/1111.6453>, 2013
- [5] V.S. Borkar, J. Nair, N. Sanketh, *Manufacturing Consent*, 48th Annual Allerton Conference, Allerton House, UIUC, Illinois, September 2010, pp. 1550-1555
- [6] V. Borkar, P. Varaiya, *Asymptotic Agreement in Distributed Estimation*, Transactions on Automatic Control, Vol. AC-27, No. 3, pp. 650-655
- [7] J. Bilmes, *Submodularity in Machine Learning Applications*, Twenty Ninth Conference on Artificial Intelligence, AAAI-I5 Tutorial Forum, January 2015
- [8] A. Björner, G. Ziegler, *Introduction to Greedoids*, in Matroid Applications (ed. N. White), Encyclopedia of Mathematics, Vol. 40, Cambridge University Press, London, UK, 1992, pp.284-357
- [9] S. Boyd, A. Ghosh, B. Prabhakar, D. Shah, *Randomized Gossip Algorithms*, IEEE Transactions on Information Theory, Vol. 52, No. 6, pp.2508-2530, June 2006

- [10] G. Calinescu, C. Chekuri, M. Pal, J. Vondrak, *Maximizing a Submodular Set Function subject to a Matroid Constraint*, SIAM J. Computing, Vol. 40, No. 6, 2011, pp. 1740-1766
- [11] K.Censor-Hillel, B. Haeupler, J.Kelner, P. Maymounkow, *Global Computation in a Poorly Connected World:Fast Rumor Spreading with No Dependence on Conductance*, Proceedings of the forty-fourth Annual ACM symposium on the Theory of Computing, 2012, pp.961-970
- [12] A. Clark, L. Bushnell, R. Poovendran, *Leader Selection for Minimizing Convergence Error in Leader-Follower Systems:A Supermodular Optimization Approach*, 10th International Symposium Modeling and Optimization in Mobile, Ad-Hoc and Wireless Networks (WiOpt), May 2012, pp. 111-115
- [13] A.Clark, B. Alomair, L. Bushnell, R.Poovendran, *Submodularity in Dynamics and Control of Networked Systems*, Communications and Control Engineering, Springer International Publishing 2016
- [14] T.H. Cormen, C.E. Leiserson, R.L. Rivest, C. Stein *Introduction to Algorithms*, third edition, MIT Press, 2009
- [15] U.Feige, *A threshold of $\ln n$ for approximating set cover* J. ACM 45 (1998) 634-652
- [16] S. Fujishige, *Submodular Functions and Optimization*, Springer Verlag, Heidelberg, London, New York, Singapore, Tokyo, 2005
- [17] G.Giakoulopis, *Tight Bounds for Rumor Spreading with Vertex Expansion*, Proceedings of the 25th ACM-SIAM Symposium on Discrete Algorithms (SODA), 2014, pp.801-815
- [18] F. Hunt, *The Structure of Optimal and Near Optimal Sets in Consensus Models*, NIST Special Publication 500-303, August 2014, <http://arxiv.org/abs/1408.4364> (online version)
- [19] F.Hunt, *An Algorithm for Identifying Optimal Spreaders in a Random Walk Model of Network Communication*, Journal of Research of the National Institute of Standards and Technology, Volume 121 (2016) <http://dx.doi.org/10.6028/jres.121.008>
- [20] V. Ilev, *An approximation guarantee of the greedy descent algorithm for minimizing a supermodular set function*, Discr. Appl. Math., **114**, pp. 131-146, 2001
- [21] R.Iyer, S.Jegelka, J.Bilmes, *Fast Semi-Differential Based Submodular Based Optimization*(longer version) arXiv:1308.1006v1 [cs.DS] 5 Aug 2013, Proceedings of the 30th International Conference on Machine Learning, Atlanta, Georgia, USA, 2013
- [22] R.Iyer, S. Jegelka, J. Bilmes, *Polyhedral aspects of Submodularity, Convexity and Concavity* arXiv:1506:07329v2, September 2015
- [23] A. Jadbabaie, *On geographic routing without location information*, 43rd IEEE Conference on Decision and Control, Vol. 5, pp.4764-5769
- [24] D. Jungnickel, *Graphs, Networks, and Algorithms*, Springer Verlag, New York, Berlin, Heidelberg, Tokyo, 1991
- [25] J. Kemeny, J. Snell, *Finite Markov Chains*, 2nd edition, Springer-Verlag, New York, Berlin, Heidelberg, Tokyo, 1976

- [26] D. Kempe, J. Kleinberg, E. Tardos, *Maximizing the Spread of Influence through a Social Network*, Proceedings of the 9th ACM-SIGKDD International Conference, Washington D.C. 2003, pp.137-146
- [27] S.Khuller, A.Moss, J. Naor, *The Budgeted Maximum Coverage Problem*, Inform. Process Lett. 70 (1999), 39-45
- [28] B. Korte, L. Lovasz, R. Schrader, *Greedoids*, Algorithms and Combinatorics Series, Vol. 4, Springer Verlag, Berlin, Germany, 1991
- [29] A. Krause, J. Leskovec, C. Guestrin, J. VanBriesen, C. Faloutsos, *Efficient Sensor Placement Optimization for Securing Water Distribution Networks*, Journal of Water Resource Planning and Management, Vol. 134, No. 6, Nov. 1, 2008
- [30] D. Mikesell, F.H.J. Kenter, I.V. Hicks, *Finding optimal hitting sets for random walks on networks*, preprint, 2015
- [31] R. Lambiotte, R. Sinatra, J.C.Delvenne, T.S. Evans, M. Barahona, V.Lattora, *Interweaving dynamics and structure*, Phys Rev. E **84**, 017102, 2011
- [32] Grimmett, G., Stirzaker, D., *Probability and Random Processes*, 3rd edition, Oxford University Press Inc., Oxford, New York, 2001
- [33] G.L. Nemhauser, L.A. Wolsey, *Best algorithms for approximating the maximum of a submodular set function*, Math. Oper. Res. 3, 1978, pp. 177-188
- [34] G.L. Nemhauser, L.A. Wolsey, M.L. Fisher, *An analysis of approximations for maximizing submodular set functions-I*, Mathematical Programming, Vol. 14, 1978, pp. 265-294,
- [35] R. Olfati-Saber, J. Fax, R.M. Murray, *Consensus and cooperation in networked multi agent systems*, Proceedings IEEE, Vol. 95, January 2007, pp. 215-233
- [36] M. Pirani, S. Sundaram, *Spectral Properties of the Grounded Laplacian Matrix with Applications to Consensus in the Presence of Stubborn Agents*, Proceedings of ACC2014, the 33rd, American Control Conference, Portland OR, 2014
- [37] A.Rao, S.Ratnasamy, C. Papadimitriou, S. Shenker, I. Stoica, *Geographic Routing without Location Information*, Proceedings of the 9th annual international conference on Mobile computing and networking, 2003, pp. 96-108
- [38] M.Richardson, P.Domingos, *Mining Knowledge Sharing Sites for Viral Marketing*, Eighth International Conference on Knowledge, Discovery and Data Mining, 2002
- [39] S. Sahni, *Approximate algorithms for the 0/1 knapsack problem*, J. ACM 22 (1975), 115-124
- [40] M. Sviridenko, *A note on maximizing a submodular set function subject to a knapsack constraint*, Operations Research Letters, 32, (2004), 41-42
- [41] M.Sviridenko, J. Vondrak, J. Ward, *Optimal Approximation for submodular and supermodular with bounded curvature*, accepted SODA15, arXiv:1311.4728v3, December 2014
- [42] J. Tsitsiklis, PhD Thesis, Massachusetts Institute of Technology, 1984
- [43] Z.Wang, B.Moran, X.Wang, Q.Pan, *Approximation for maximizing monotone non-decreasing set function with a greedy method*, J.Comb. Optimization (online) DOI 10.1007/s10878-014-9707-3, January 2014